Phylogeny

Codon models

- Last lecture: poor man's way of calculating dN/dS (Ka/Ks)
  - Tabulate synonymous/non-synonymous substitutions
  - Normalize by the possibilities
  - Transform to genetic distance K_{JC} or K_{k2p}
- In reality we use codon model
  - Amino acid substitution rates meet nucleotide models
  - Codon(nucleotide triplet)

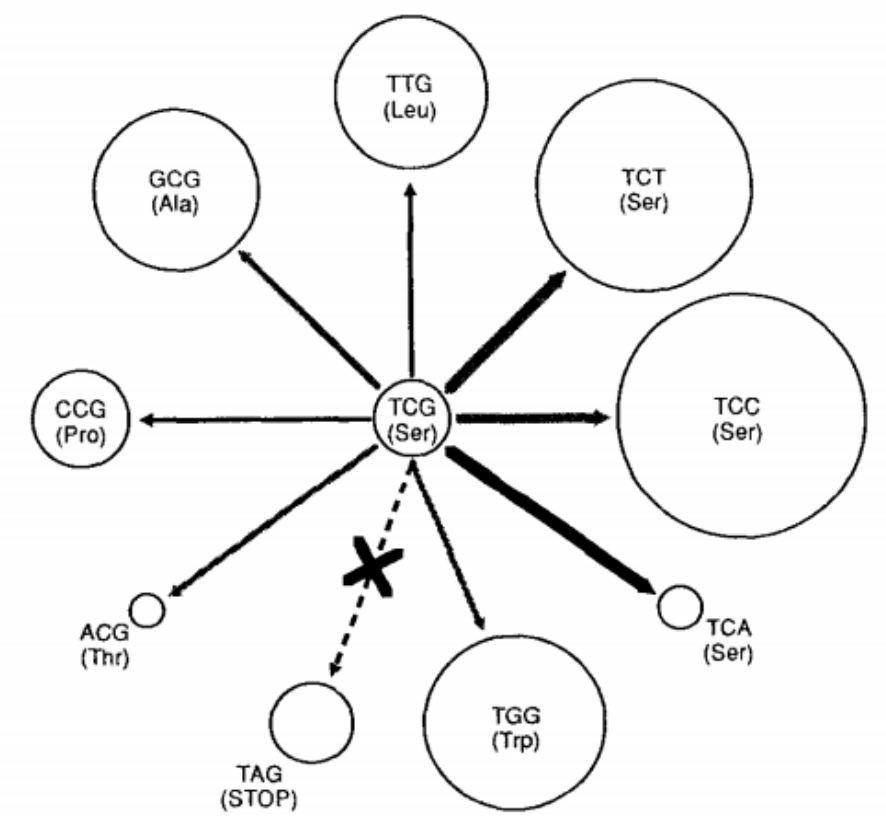


FIG. 1.—Example of the "neighbors" to which a codon (here, TCG) may evolve instantaneously through a nucleotide substitution at one position. TCG has eight neighbors, substitution of A for C at the second position being disallowed, as it results in the stop codon TAG. Transitions are marked with black arrows, transversions with gray arrows. Substitutions involving no change in amino acid (generally occurring at a higher rate in this model) are marked with thicker arrows. The size of each circle (except the stop codon TAG) represents the (equilibrium) frequency of that codon, in this case taken from the pooled α- and β-globin gene sequences.

Codon model parameterization

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases}$$

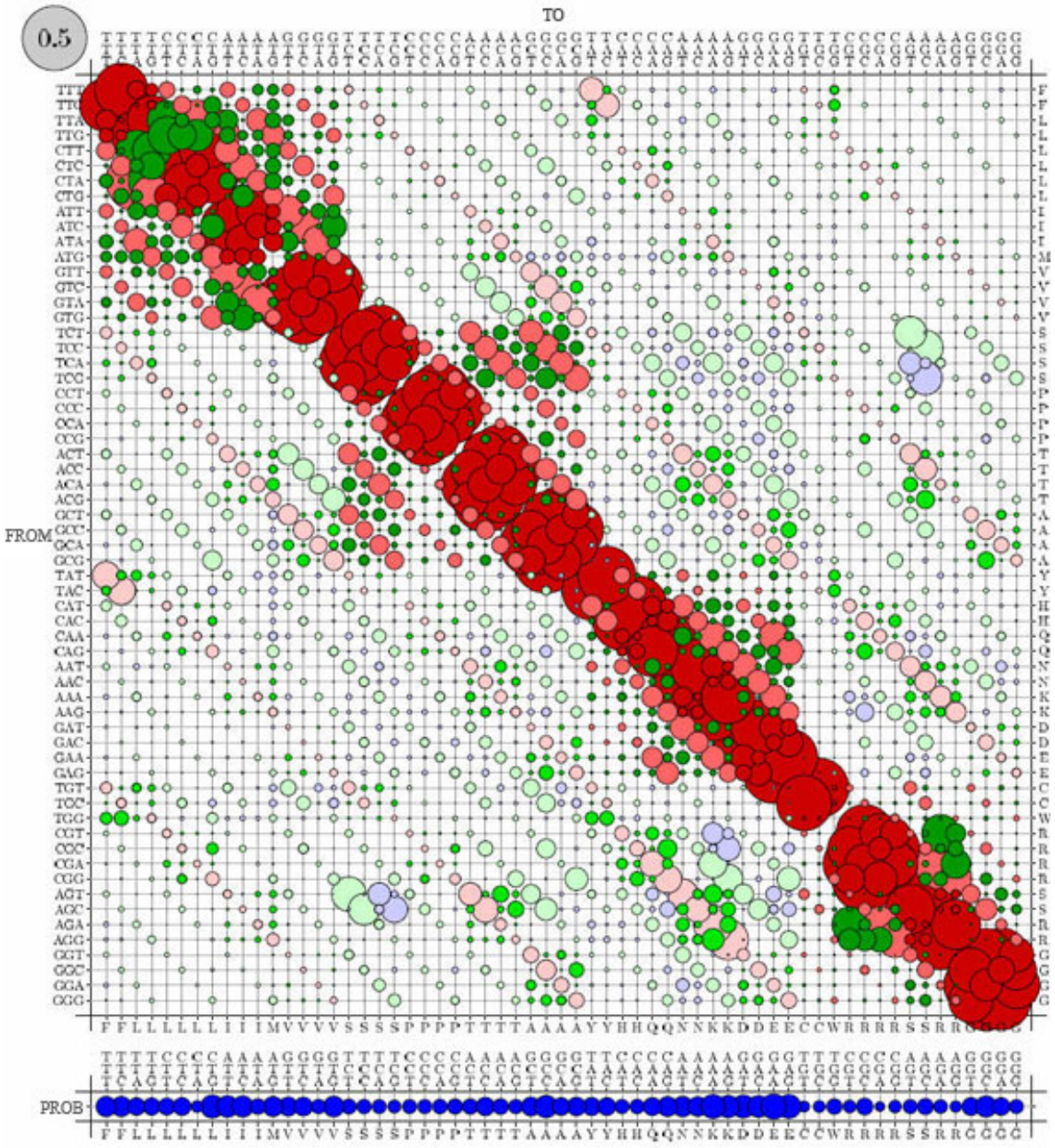
Stop codons are not allowed, reducing the matrix from 64x64 to 61x61

The entire codon matrix can be parameterized using:

- κ kappa, the transition/transversion ratio
- ω omega, the d_N/d_S ratio – optimizing this parameter gives the an estimate of selection force
- π_j the equilibrium codon frequency of codon j

(Goldman and Yang. MBE 1994)

Empirical codon substitution matrix



Observations:

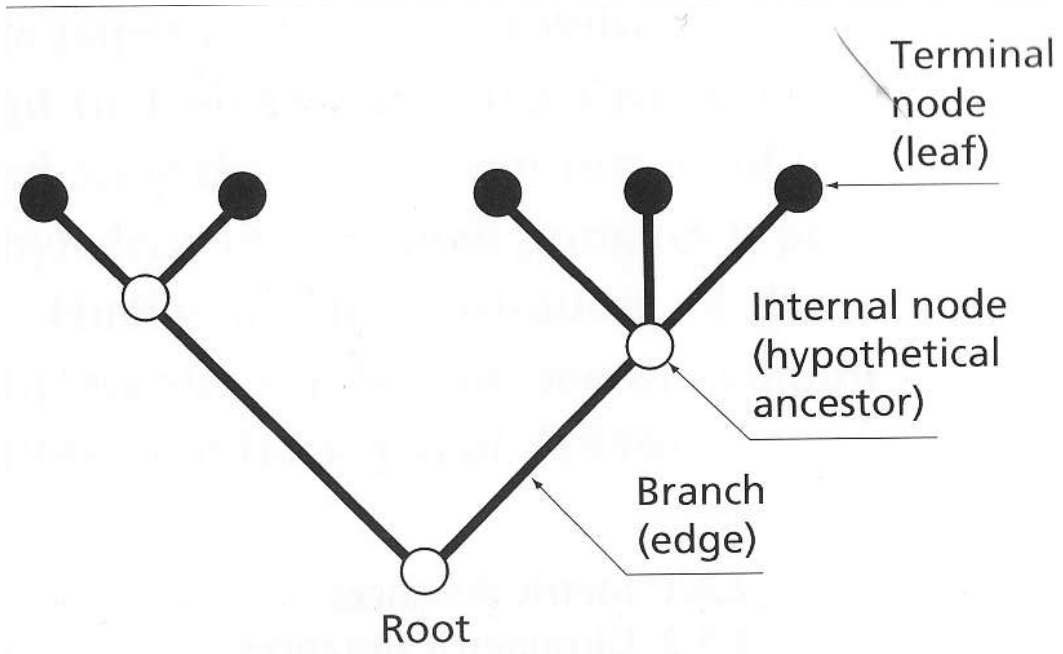
Instantaneous rates of double nucleotide changes seem to be non-zero

There should be a mechanism for mutating 2 adjacent nucleotides at once!

(Kosiol and Goldman)

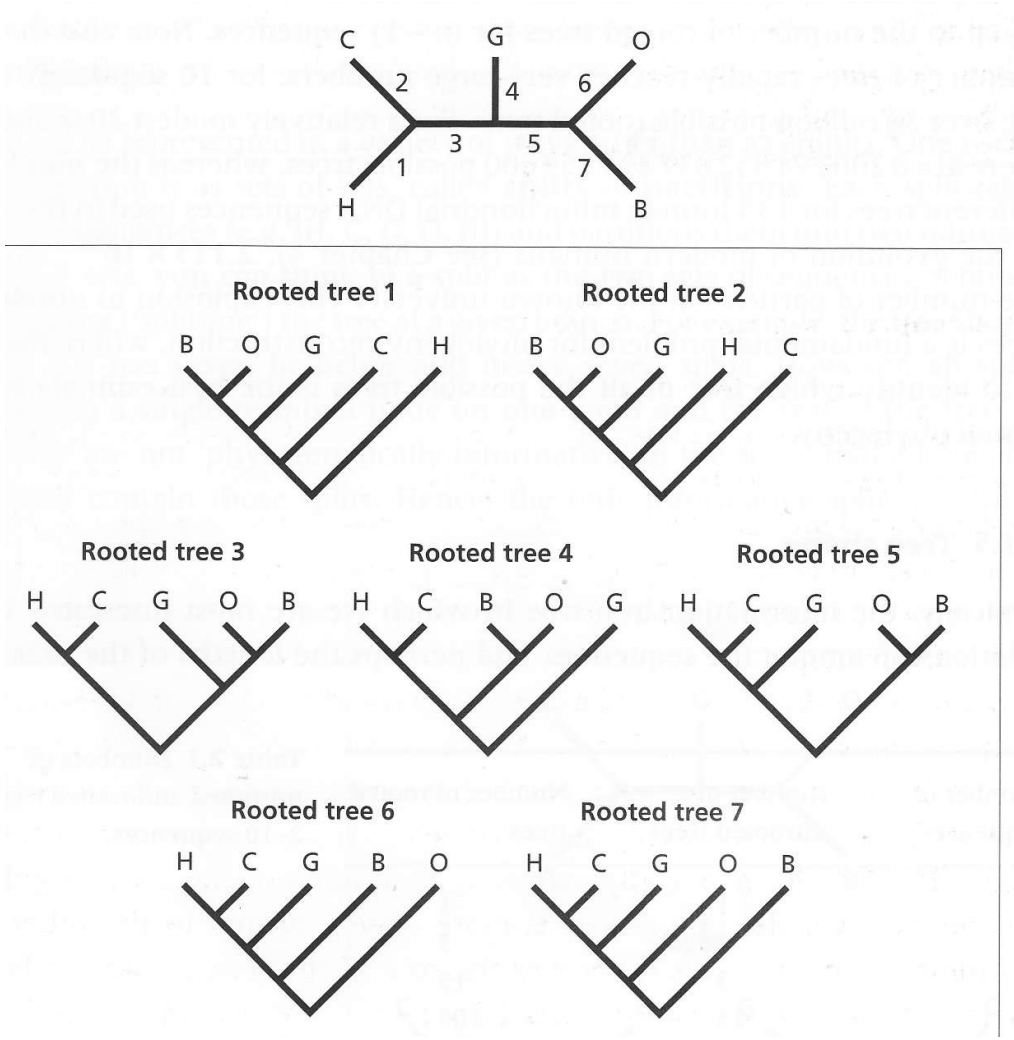
Phylogeny

- Phylogenetic trees
  - Topology
  - Branch length
- Last lecture: Inferring distance from an alignment
- How do we infer trees given distance
- How to infer trees and distance given an alignment



Rooted vs unrooted

- What is a good tree
  - Accurately represents observed distance
  - Distance goodness of fit
  - Accurately represents evolutionary history—mode difficult to define
    - Depends on assumption about evolutionary process
- A tree root implies a common ancestor
  - Some tree building approaches do not provide a root
  - Root identification does effect goodness of fit



Rooted

N species

2N – 2 branches

N – 1 internal nodes

Unrooted

N species

2N – 3 branches

N – 2 internal nodes

Gene tree vs species tree

- Gene tree is not always the same as a species tree
- Species evolve as populations--gene divergence can predate a speciation event
- Inferring species relationship from molecular data—use many orthologous genes –paralogs present many problems
- Orthologs can be identified using synteny—chromosomal order

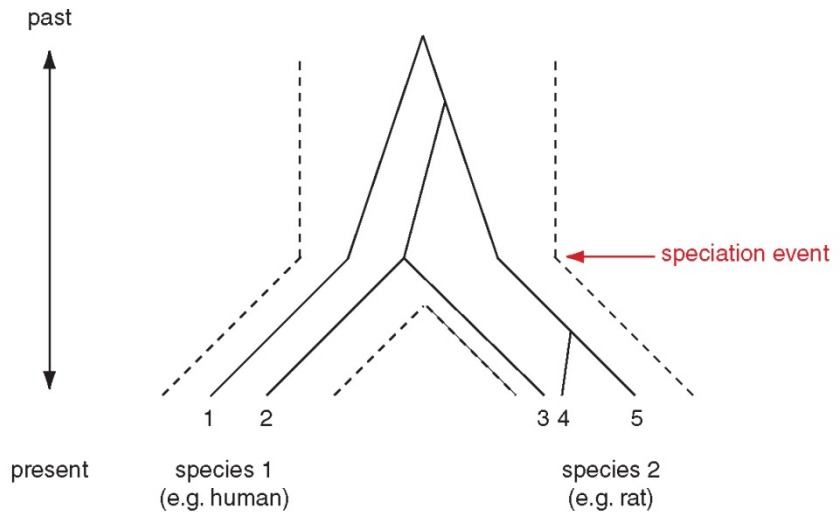


FIGURE 7.13 A species tree and a protein (or gene) tree can have a complex relationship. A speciation event, such as the divergence of the lineage that generated modern humans and rodents, can be dated to a specific time (e.g., 90 MYA). When speciation occurs, the species become reproductively isolated from one another. This event is represented by dotted lines (see horizontal arrow). Phylogenetic analysis of a specific group of homologous proteins is complicated by the fact that a gene duplication could have preceded or followed the speciation event. In essentially all phylogenetic analyses, the extant proteins (OTUs) are sequences from organisms that are alive today. It is necessary to reconstruct the history of the protein family as well as the history of each species. In the above example, there are two human paralogs and three rat paralogs. Proteins 1 and 5 diverged at a time that greatly predates the divergence of the two species. Proteins 2 and 3 diverged at a time that matches the date of species divergence. Proteins 4 and 5 diverged recently, after the time of species divergence. It is possible to reconstruct both species trees and protein (or gene) trees. Adapted from Graur and Li (2000), based upon Nei (1987). Reproduced with permission from Sinauer Associates and Columbia University Press.

Phylogeny is reconstructed through either distance data or discrete characters

Distance – all pairwise distances are computed from characters.

- UPGMA, Neighbor-Joining…

	hg19	panTro2	gorGor1	ponAbe2	rheMac2	papHam1	calJac1
hg19	0.000000	0.012640	0.017859	0.015237	0.052272	0.049648	0.063510
panTro2	0.012640	0.000000	0.020448	0.017809	0.054991	0.052361	0.066286
gorGor1	0.017859	0.020448	0.000000	0.023110	0.047165	0.049920	0.066730
ponAbe2	0.015237	0.017809	0.023110	0.000000	0.052298	0.049673	0.057901
rheMac2	0.052272	0.054991	0.047165	0.052298	0.000000	0.007565	0.080637
papHam1	0.049648	0.052361	0.049920	0.049673	0.007565	0.000000	0.083558
calJac1	0.063510	0.066286	0.066730	0.057901	0.080637	0.083558	0.000000

Discrete character methods use sequences directly during inference.

- Maximum parsimony, Maximum likelihood, Bayesian methods…

```
 10  399
hg19       ATGGGATCTTCTGGACTTTTGAGCCTCCTGGTGC
panTro2    ATGGGATCTTCTGGACTTTTGAGCCTCCTGGTGC
gorGor1    ATGGGATCTTCTGGACTTTTGAGCCTCCTGGTGC
ponAbe2    ATGGGATCTTCTGGACTTTTGAACCTCCTGGTGC
rheMac2    ATGGGATCTTCTGGACTTTTGAGCCTCCTGGTGC
papHam1    ATGGGATCTTCAGGACTTTTGAGCCTCCTGGTGC
calJac1    ATGGGATCTTCTGGACTTTTGAGCCTCTTGGTGC
tarSyr1    ATGGAATCATCTAAACTTTTGAGCCTCCTGGTGC
micMur1    ATGGAATATTCCGGACTTTTGAGCCTCCTGGTGC
tupBel1    ATGGAATCTTCTGGACTTCTGAGCATCGTGGTGT
```

UPGMA- unweighted Pair Group Method with Arithmetic Mean

First, join shortest-distance pair with branching point at half distance.

$$L_{ij} = d_{ij} / 2$$

Recalculate distance matrix joining x&y as node1 (n1).

$$L_{n1z} = (d_{xz} + d_{yz} - d_{xy}) / 2$$

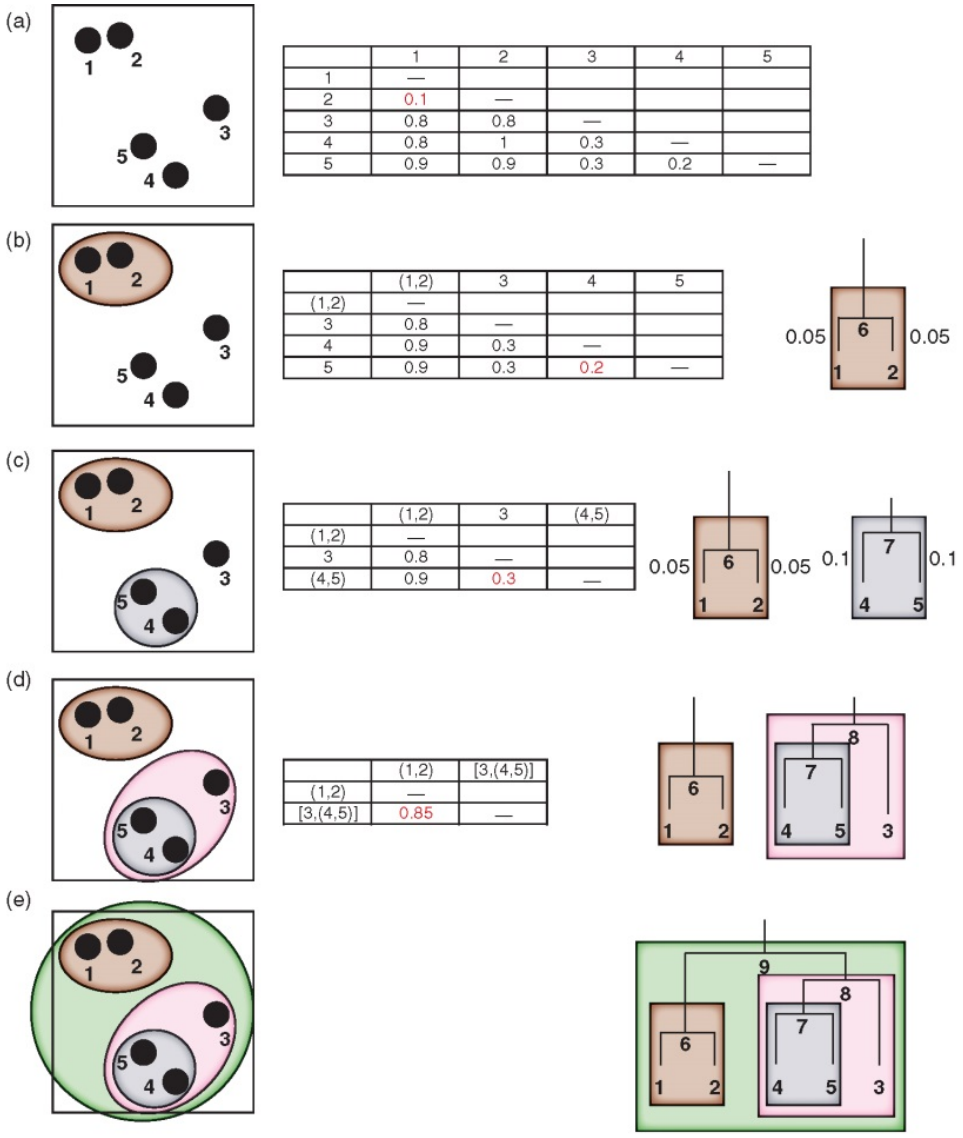
Join next shortest-distance pair, and continue…

When updating distances to clusters, use the size of the cluster to calculate correct distance

$$d_{(\mathcal{A}\cup\mathcal{B}),X} = \frac{|\mathcal{A}|\cdot d_{\mathcal{A},X} + |\mathcal{B}|\cdot d_{\mathcal{B},X}}{|\mathcal{A}|+|\mathcal{B}|}$$

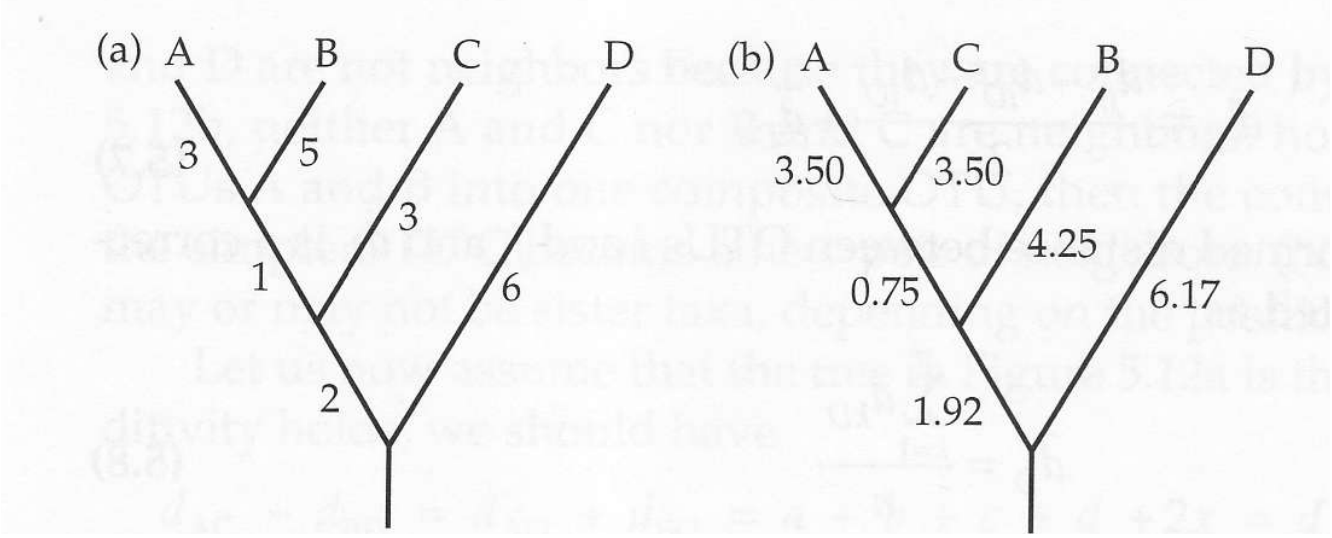
If genetic distance perfectly correlated with time (molecular clock), the UPGMA method would return the correct tree topology and branch lengths.

This is never true, and UPGMA, although simple and useful for learning trees, is not often used.



UPGMA mistakes

- UPGMA is not often used and makes mistakes by joining branches with slow rates, but that are not true sister taxa.
- produces an ultrametric tree in which the distances from the root to every branch tip are equal



	A	B	C
B	8	7	12
C	7	9	14
D	12	14	11

Neighbor Joining

- Greedy like UPGMA –except…
- Start with everything in a star topology and gradually join leaves with a common branch until a binary tree remains
- Which leaves to join
  - Transform distance matrix into a U matrix

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

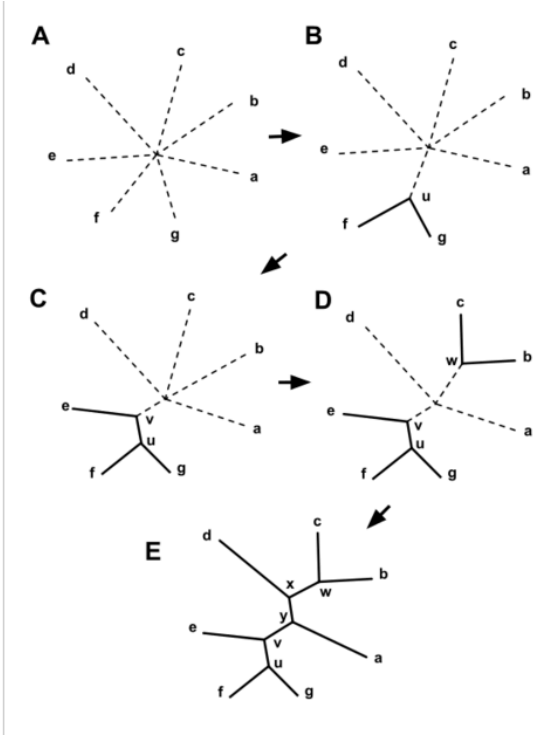
where $d(i,j)$ is the distance between taxa i and j.

- Calculate new distance: we joined g and f to create internal node u

$$\delta(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)\right]$$

$$\delta(g,u) = d(f,g) - \delta(f,u)$$

- Each calculations uses information from the entire distance matrix



D

	a	b	c	d	e
a	0	5	9	9	8
b	5	0	10	10	9
c	9	10	0	8	7
d	9	10	8	0	3
e	8	9	7	3	0

31
34
34
30
27

U

	a	b	c	d	e
a		–50	–38	–34	–34
b	–50		–38	–34	–34
c	–38	–38		–40	–40
d	–34	–34	–40		–48
e	–34	–34	–40	–48	

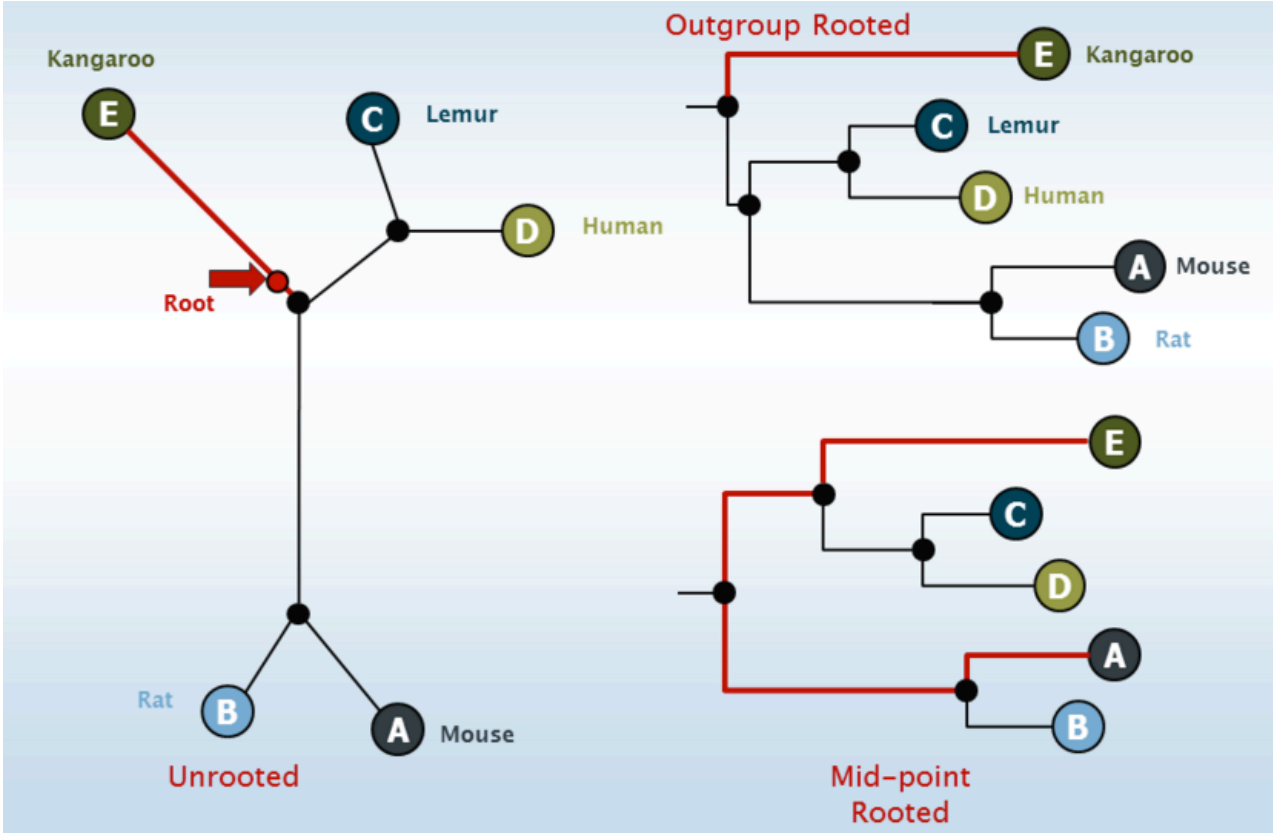
Figures and formulas from Wikipedia

NJ properties

- NJ can be seen as optimizing the least squares error to the distance matrix
- Neighbor joining has the property that if the input distance matrix is correct, then the output tree will be correct.
- The correctness of the output tree topology is guaranteed as long as the distance matrix is 'nearly additive'-- if each entry in the distance matrix differs from the true distance by less than half of the shortest branch length in the tree
- In practice above is not satisfied but the correct tree is found anyway
- May assign negative lengths
- Tree has no root

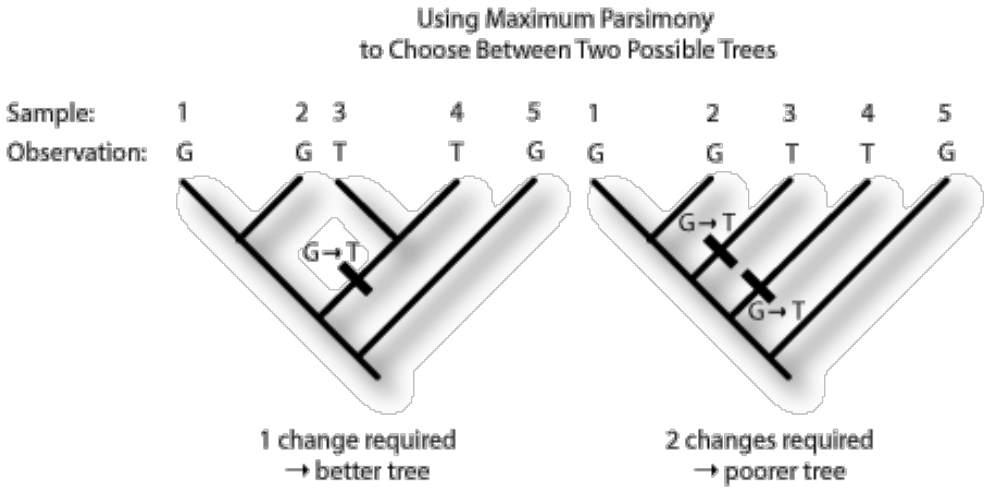
Rooting a tree

- Midpoint rooting
  - Find the longest path and place the root in the middle
  - works if rate is constant or balanced along the tree
  - Shouldn't change if species is removed
- Outgroup rooting-using prior knowledge about relationships



Maximum Parsimony

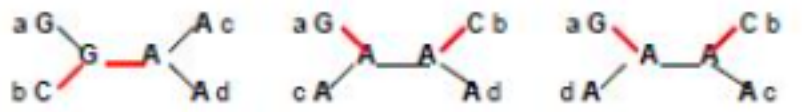
- Not greedy—explores the tree space and uses sequence alignments directly
- Uses an **optimality** criterion to scrutinize trees. Therefore, it is a tree searching method, as opposed to clustering like distance-based methods.
- Criterion – smallest number of evolutionary changes
  - Ockham's razor: the best hypothesis is that requiring the fewest assumptions.



Maximum parsimony

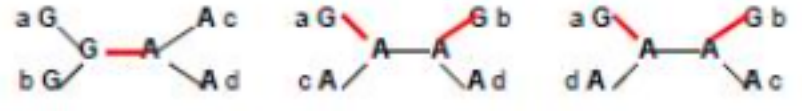
	1	2	3	4	5	6	7	8	9
a	A	A	G	A	G	T	T	C	A
b	A	G	C	C	G	T	T	C	T
c	A	G	A	T	A	T	C	C	A
d	A	G	A	G	A	T	C	C	T

site 3

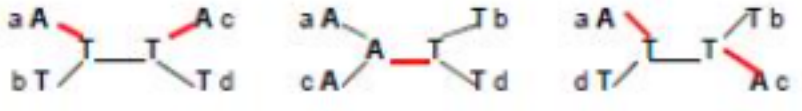


Same as Site 7

site 5



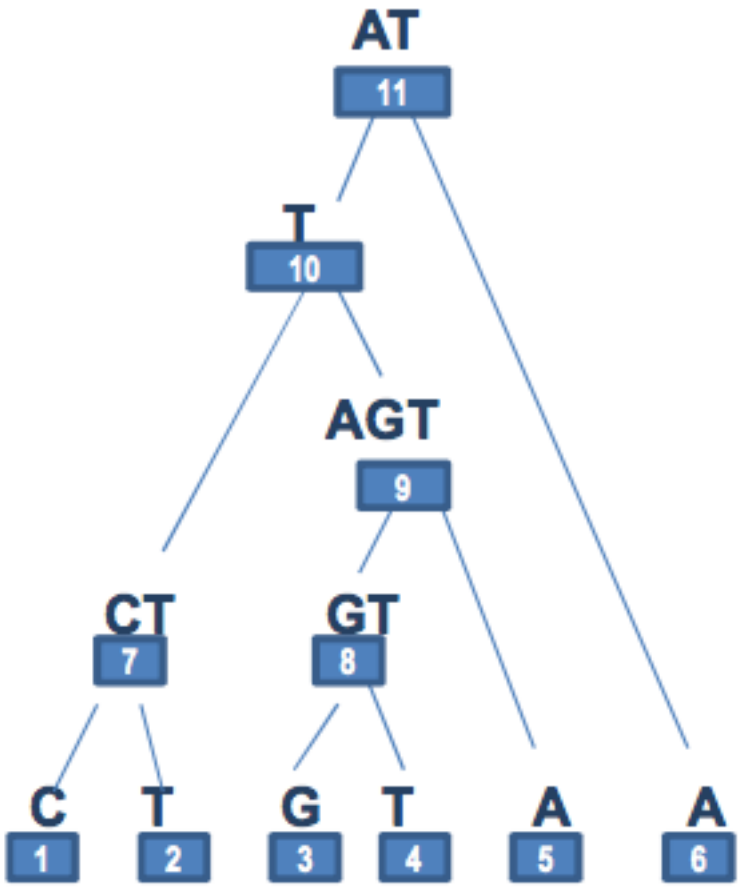
site 9



- Sites are informative if
  - they are variable
  - The help us differentiate between trees
  - At least 2 characters occur at least 2 times
- Site2 is uninformative because all three possible trees require 1 evolutionary change, G ->A.
- Site 3 is uninformative because all trees require 2 changes.
- Site4 is uninformative because all trees require 3 changes.
- Site5 is informative because tree I requires one change, trees II and III require two changes
- Site7 is informative,likesite 5
- Site 9 is informative because tree II requires one change, trees I and III require two.

site9+2*site5 4 5 6

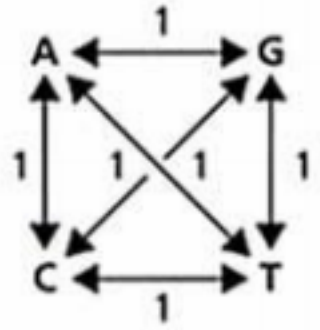
Parsimony rule



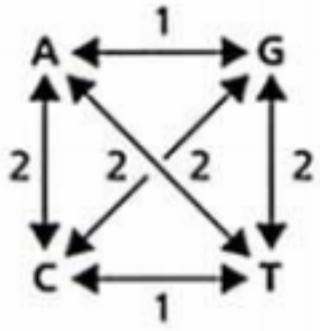
- The set at an interior node is the intersection of its two immediately descendant sets if the intersection is not empty.
- Otherwise it is the union of the descendant sets.
- For every occasion that a union is required to form the nodal set, a nucleotide substitution at this position must have occurred at some point during the evolution for this position.
- Number of union operations = minimum number of substitutions required to account for descendant nucleotides

Parsimony in practice

- Weighted parsimony
- Changes have different cost
  - Transitions/transversions



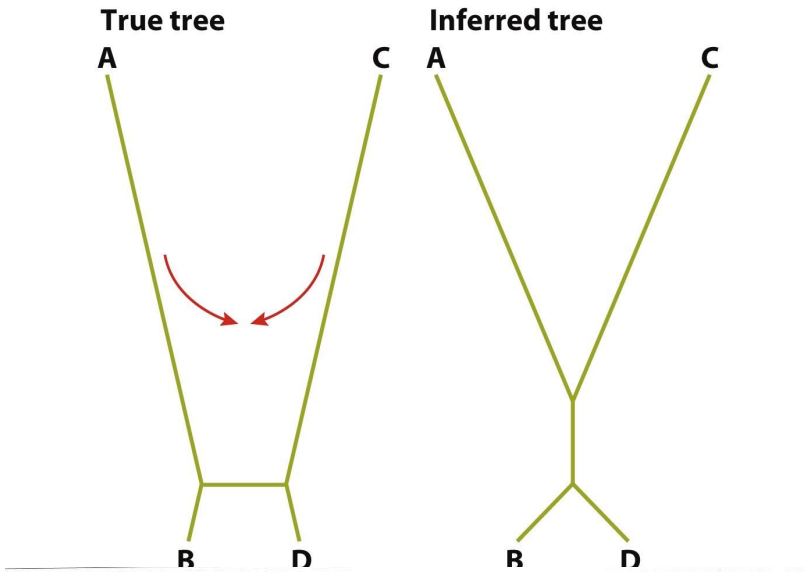
From \ To	A	C	G	T
A	0	1	1	1
C	1	0	1	1
G	1	1	0	1
T	1	1	1	0

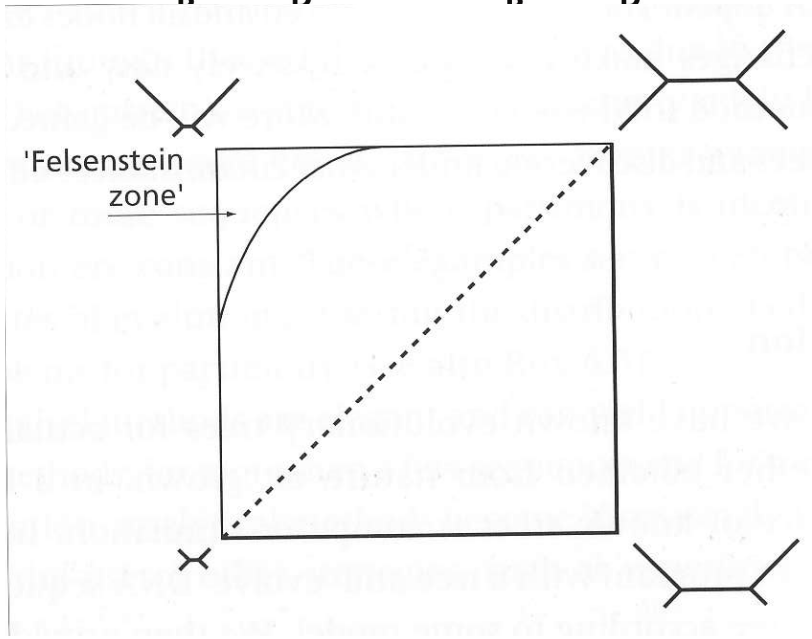


From \ To	A	C	G	T
A	0	2	1	2
C	2	0	2	1
G	1	2	0	2
T	2	1	2	0

Maximum parsimony--Long branch attraction

- Tends to merge long branches together
- Felsenstein 1978





Tree search methods

- "Exhaustive" searching is possible for small numbers of species, and guarantees the optimal tree(s).

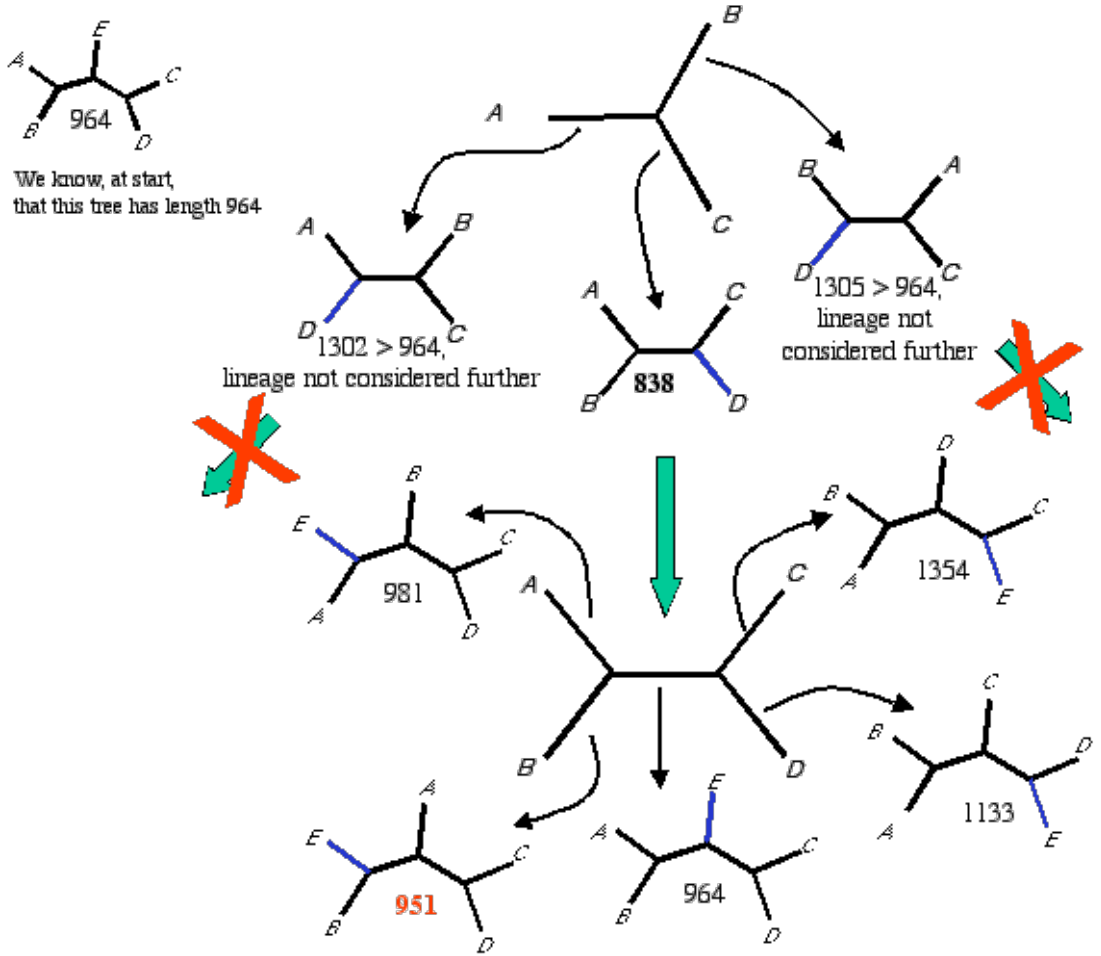
- The branch-and-bound algorithm will also search all trees but much more efficiently
  - guarantees optimal tree for up to 20 species.

- Otherwise (>20 species), heuristic methods must be applied. These are often good, but never guarantee the best tree.
- employ combinations of clustering with branch rearrangements.

n	Number of trees
1	1
2	3
3	21
4	231
5	3,495
6	67,455
7	1,584,765
8	43,897,455
9	1,400,923,755
10	50,619,052,575
11	2,042,745,514,425
12	91,066,568,444,775
13	4,444,738,893,770,175
14	235,731,740,255,186,175
15	13,499,365,993,279,291,125
16	830,161,812,269,496,081,375
17	54,564,569,247,212,367,217,875
18	3,817,309,552,613,869,238,301,375
19	283,213,212,610,863,528,421,052,625

Branch and bound

- Solution space is a tree
  - Tree of trees
- Observation: adding branches can only increase tree distance
- Maintain a current lower bound
  - For trees we can get a starting lower bound with some fairly good though not optimal method
- Some branches (of the solution space) can be abandoned



Getting confidence estimates on the tree

- Resample the alignment with replacement to create a dataset of the same length
- Rerun tree inference
- Specific internal nodes can be labeled with their bootstrap value
- Number of bootstraps is typically 100 or 1000

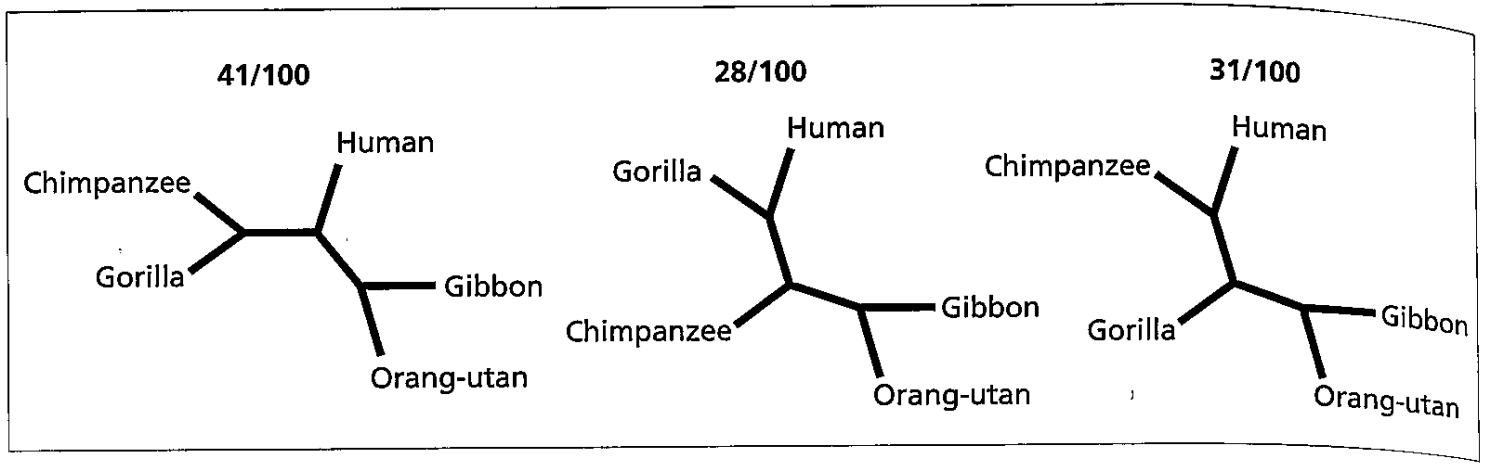
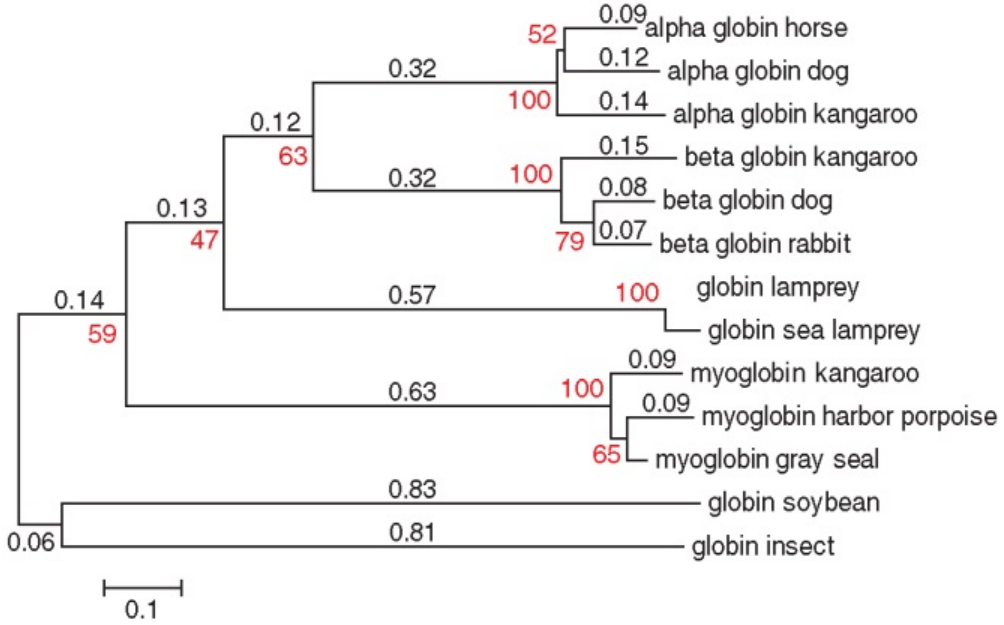
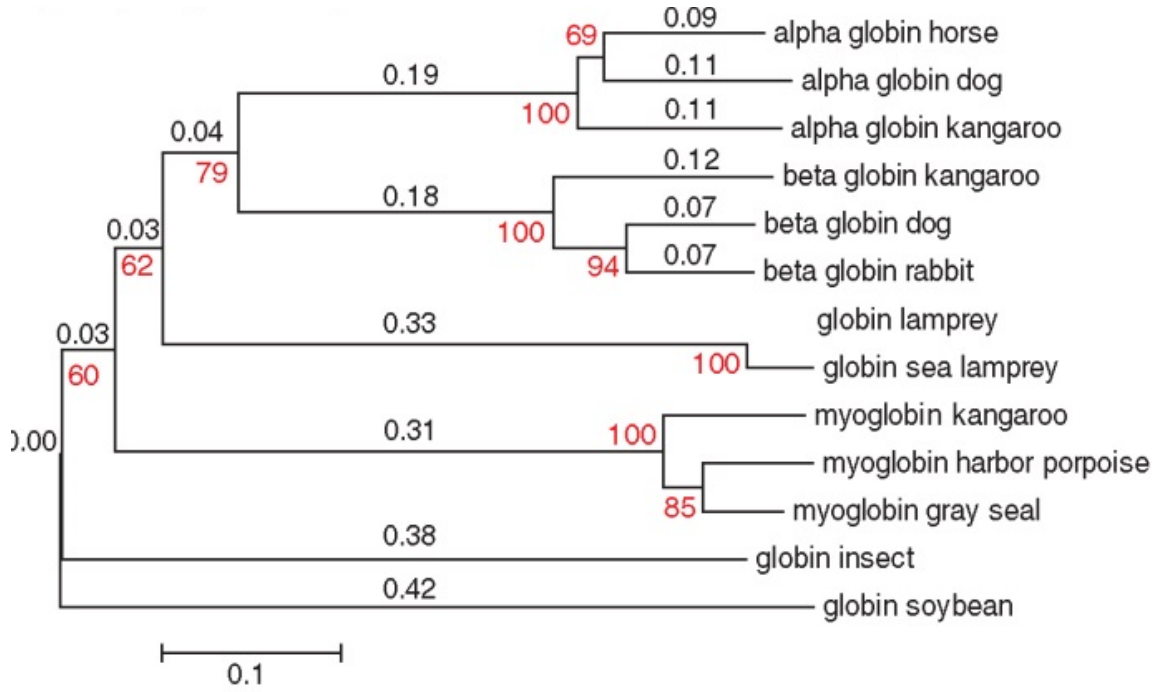


Fig. 6.37 The three trees for the hominoid sequence data that are obtained from 100 bootstrap pseudoreplicates and their relative frequencies. All three trees have the split {{orang, gibbon} {human, chimp, gorilla}} but they disagree about relationships between humans and the African apes.

Trees with bootstrap values





Maximum likelihood trees

- Most widely used method when accurate trees are required
- Use sequence directly

1. Define a model: parameter values and tree
2. Total likelihood of alignment and model is the product of likelihoods of all sites, *s,* 1 through N_s

   $L = L_1 * L_2 * \ldots * L_N = \Pi_{(s=1..N)} L_s$

   This is usually done by adding log likelihoods.
3. Likelihood of one site is the sum of all possible histories, *h*, at that site. (Histories are the states of the unknown internal nodes, *n*.)

   $L_1 = P_{1h1} + P_{1h2} + P_{1h3} + \ldots$
4. Probability of one history at one site is the product of the equilibrium frequency P_{Xeq} and all required substitution probabilities.

   $P_{1h1} = P_{Aeq} * P_{ii} * P_{ij} * \ldots$

1: Define a model: parameter values and tree

Jukes-Cantor substitution model

$P_{ii}(t) = ¼ + ¾\ e^{-4\alpha t}$ $P_{ij}(t) = ¼ - ¼\ e^{-4\alpha t}$

we want to estimate K: $K = 3\alpha t$, $\alpha t = K/3$

$P_{ii}(K) = ¼ + ¾\ e^{-4K/3}$ $P_{ij}(K) = ¼ - ¼\ e^{-4K/3}$

The tree is also part of the model: it may be given or we may want to search over all possible trees

2. Total likelihood of an alignment

Total likelihood of alignment and model is the **product** of likelihoods of all sites, *s*, 1 through N

$L = L_1 * L_2 * \ldots * L_N = \Pi_{(s=1..N)} L_s$

This is usually done by adding log likelihoods.

$\ln L = \ln L_1 + \ln L_2 + \ldots + \ln L_N = \Sigma_{(s=1..N)} \ln L_s$

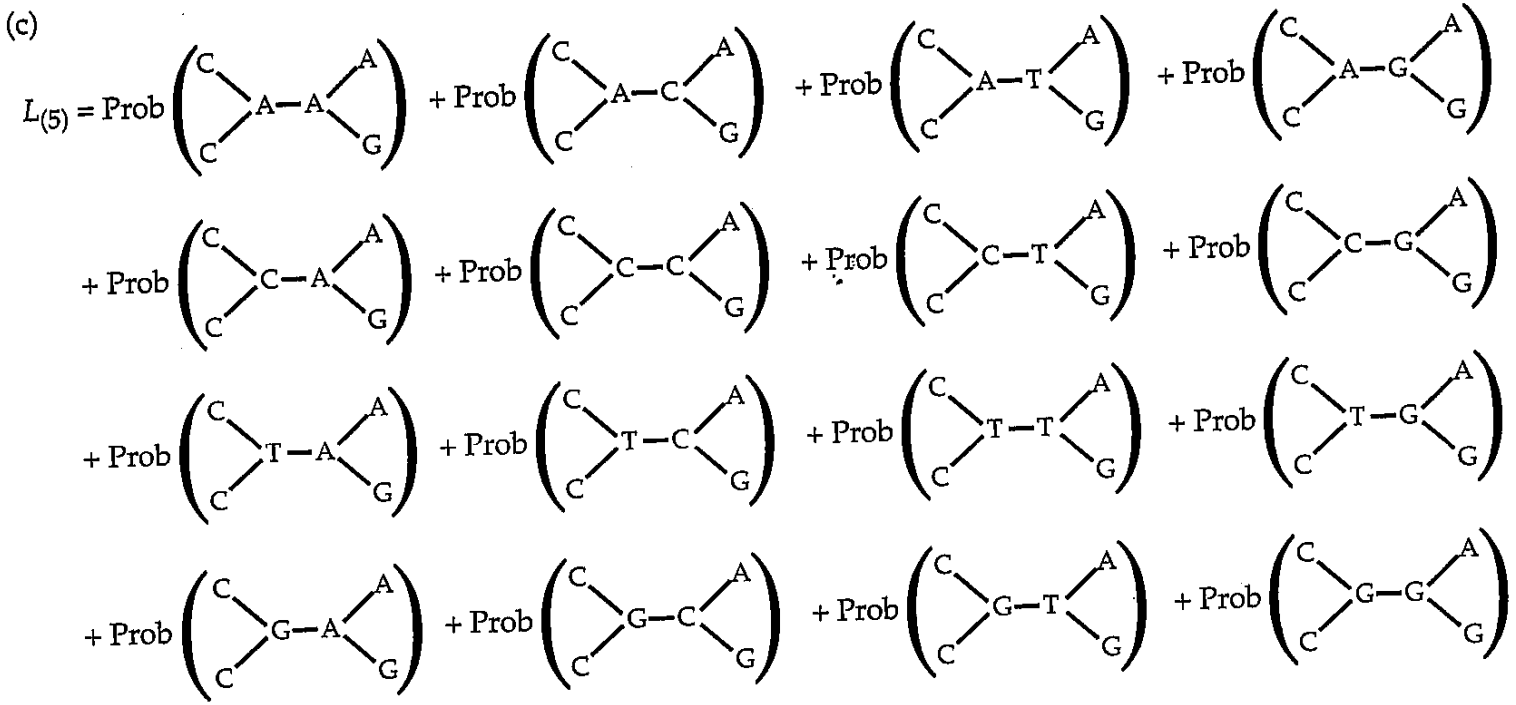
	1	2	3	4	5	6	7	8	9	...*n*
OTU1	A	A	G	A	C	T	T	C	A	...N
OTU2	A	G	C	C	C	T	T	C	T	...N
OTU3	A	G	A	T	A	T	C	C	A	...N
OTU4	A	G	A	G	G	T	C	C	T	...N

OUT: Operational Taxonomic Units.

3: Likelihood of one site

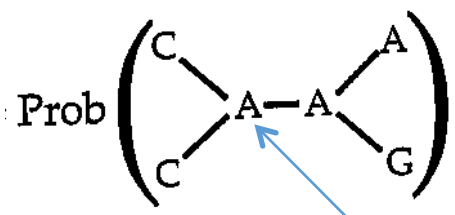
the **sum** of all possible histories, *h*, at that site. (Histories are the states of the unknown internal nodes.)

$L_1 = P_{1h1} + P_{1h2} + P_{1h3} + \ldots$ There are $4^{(T-2)}$ histories for T taxa



4. Probability of one history at one site

- Probability of one history at one site is the **product** of the equilibrium frequency P_{Xeq} of the root state, X, and the probabilities of all required substitutions.



$$= P_{Aeq} * P_{AC} * P_{AC} * P_{AA} * P_{AA} * P_{AG}$$

For a reversible model, the choice of root is arbitrary.

Searching the parameter space

Complicated models must be maximized through a guided trial-and-error, "hill climbing" algorithm.

1. Set initial parameter values and tree.
2. Calculate likelihood.
3. Propose new parameter value or tree.
4. Calculate likelihood.
5. Decide whether to accept the new value.
6. Repeat steps 3-5 until changes no longer improve likelihood.

local maxima can trap this algorithm below the best model, must try multiple initial parameter values.

This is computationally expensive but allows us to determine the maximum likelihood model in many cases.

PAML programs (*baseml, codeml, aaml*)

Tree likelihood example

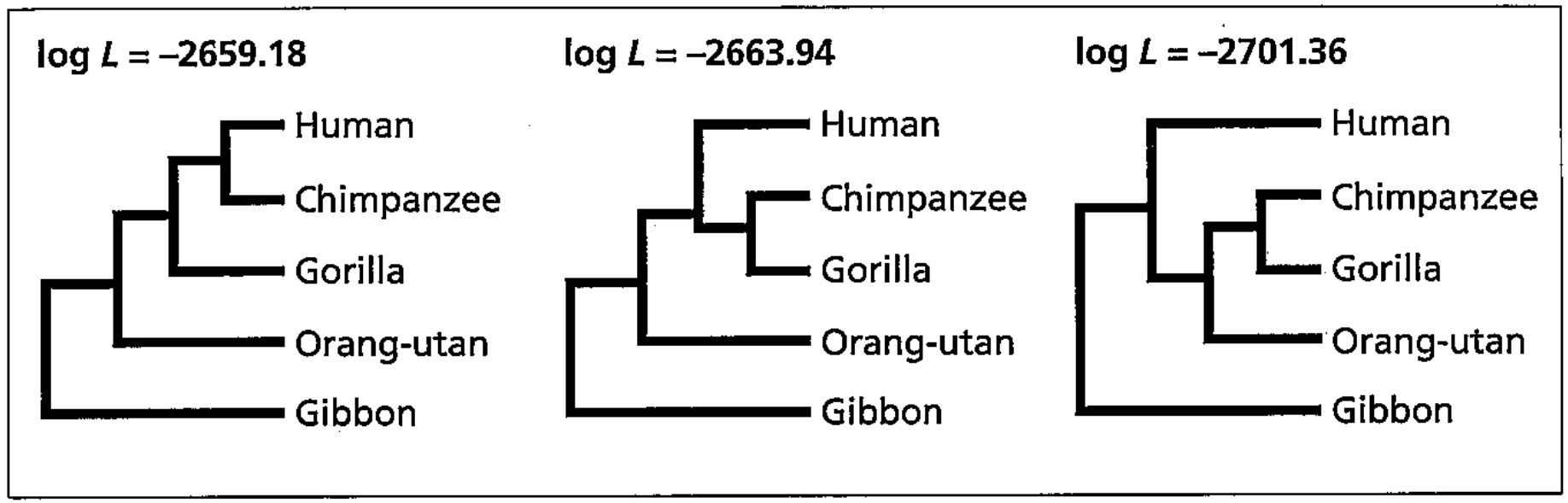
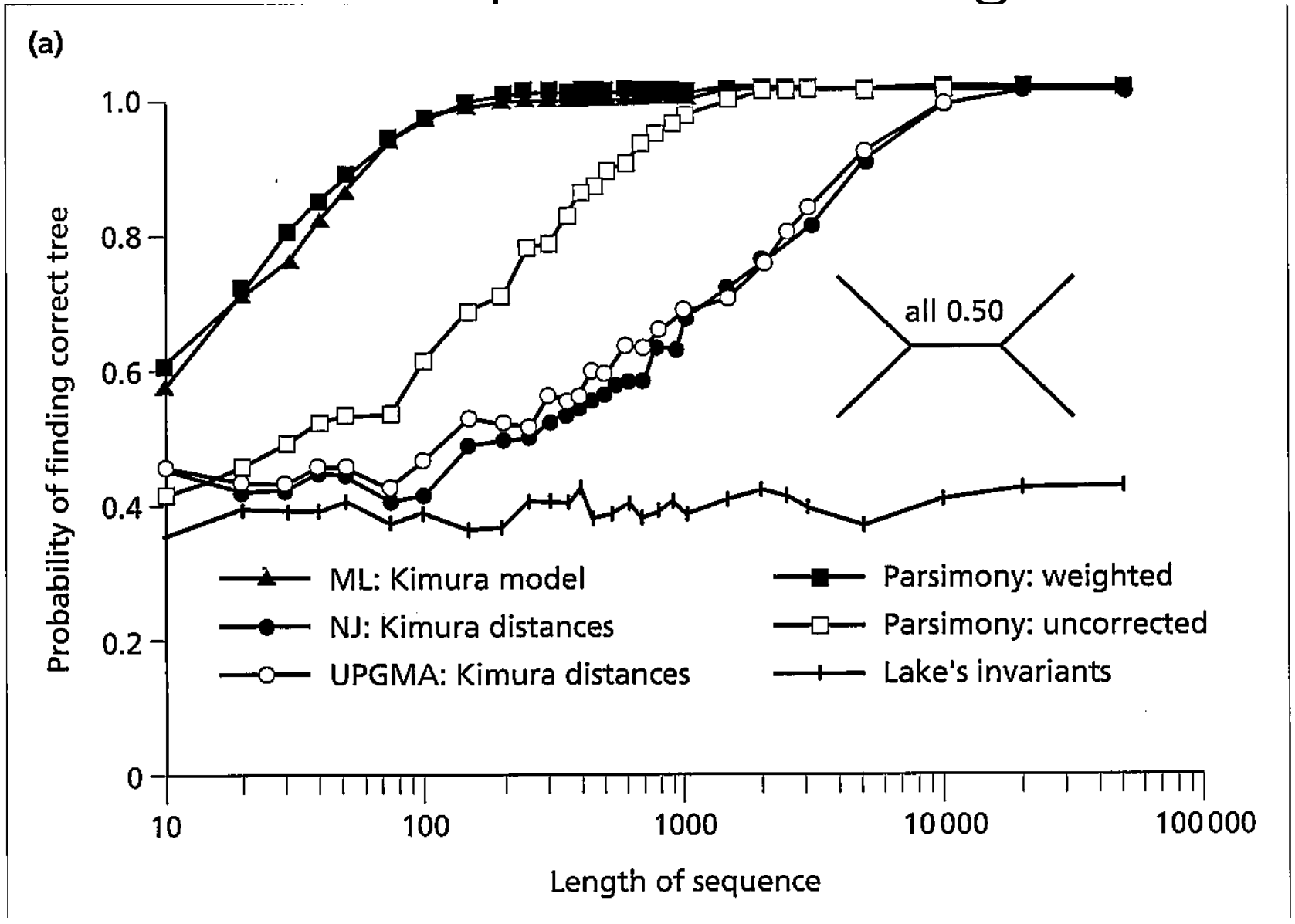
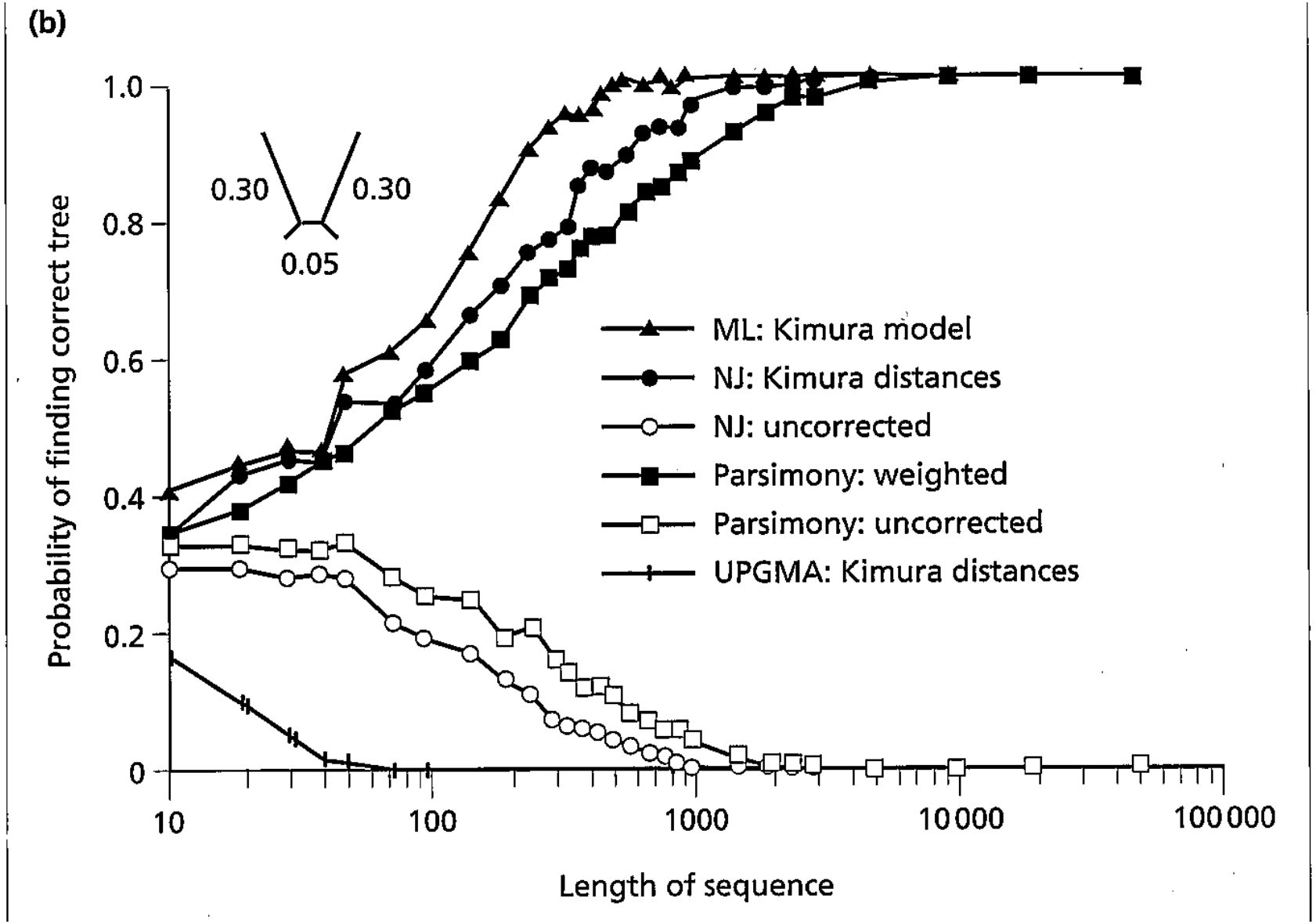


Fig. 6.19 Three different hypotheses of relationship among the hominoids and the likelihoods that each tree has given rise to the observed data.

Likelihood and weighted parsimony perform the best on trees with equal branch lengths



Likelihood performs the best on trees with **unequal** branch lengths



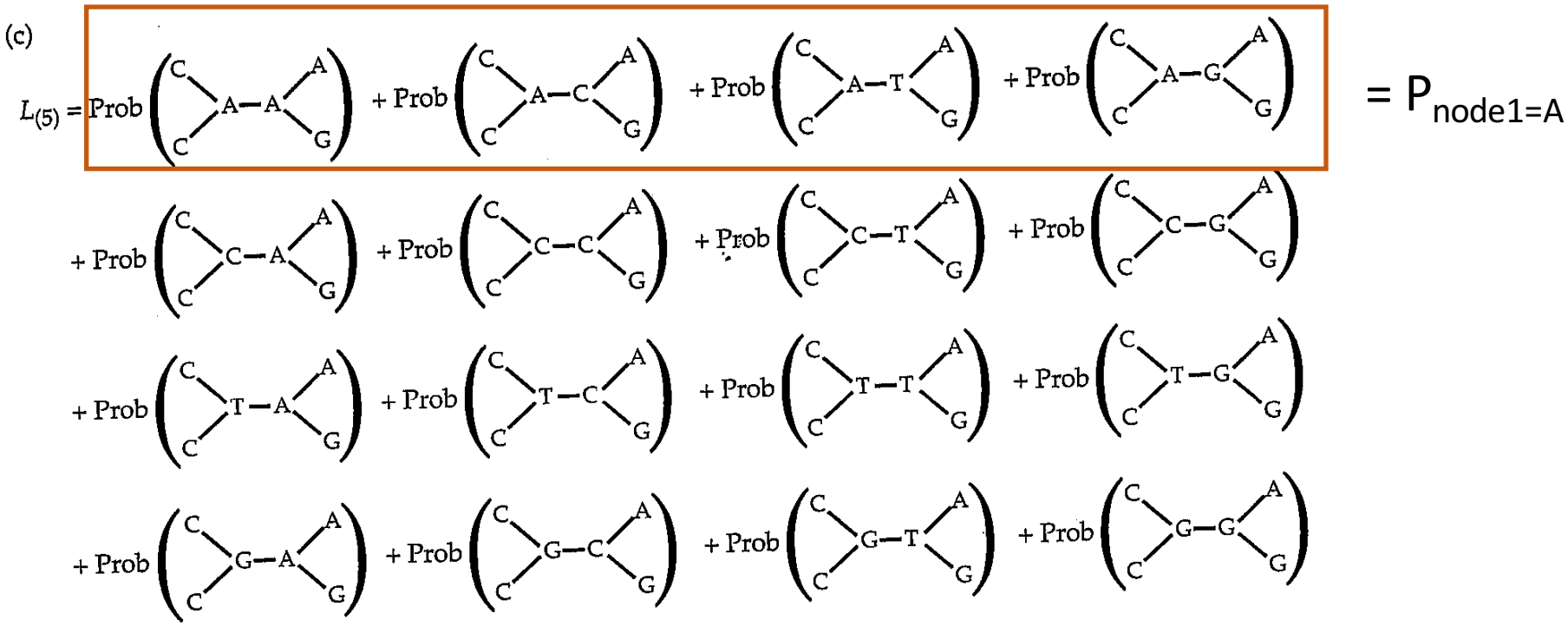
General observations about performance

- Kuhner and Felsenstein 1994
  - assumptions of independence between sites, equal rate, and unchanging model are limiting.
  - Real data or simulated data violating these assumptions posed problems for tree reconstruction.
- Yang in series of papers over 1994-1997
  - JC, K2P, and F81 models can result in severe branch length estimation errors.
  - HKY (ts/tv + equilibrium probabilities) performs better, but general time reversible (GTR) was best.

Ancestral state reconstruction

Sequences or states of last common ancestors can be inferred by summing probabilities of all histories with that state.

- This is done using the previously determined maximum likelihood tree and parameter values.



$= P_{node1=A}$

Bayesian phylogenetic

- Produces **posterior probability distributions** of parameter values, so we can **assess the range of probable values** – the credible interval (analogous to confidence interval)

  e.g. $0.31 \leq K \leq 0.46$ as a 95% credible interval of K

- MCMC – Markov Chain Monte Carlo-general method for sampling complicated distributions
  - To compute probabilities we have to integrate over many unknown parameters
  - Solution: constructing a Markov chain that has the desired distribution as its equilibrium distribution.
    - Given a current point in the parameter set define a set of rules for choosing the next one
    - Do this long enough and the distribution will converge and we can calculate statistics

- Implementation: MrBayes

Learn more

- MSCBIO 2075 Molecular Evolution
- Spring 2017
- Dennis Kostka/ Nathan Clark